DRAGON SYSTEMS' 1997 MANDARIN BROADCAST NEWS SYSTEM

Puming Zhan, Steven Wegmann, and Steve Lowe

Dragon Systems, Inc.
320 Nevada Street, Newton, MA 02160

1. INTRODUCTION

The development of our 1997 HUB4 Mandarin system was an exercise in technology transfer. For this initial implementation, our strategy was to change the structure of our HUB4 English system only when absolutely necessary. In deciding what was "necessary", we tried to bear in mind that there are differences in the languages that have important implications for speech recognition. For example, Mandarin is a toned language and our front-end does not standardly compute pitch, one of the most important indicators of tone. Also, the notion of word is not well defined for Mandarin, which has implications for language model and even acoustic model development in our word-based recognition system.

The Mandarin system we developed for the HUB4 evaluation is almost identical to our English system [1] except in the following respects:

- We used only the data supplied by the LDC specifically for the Mandarin HUB4 evaluation. In contrast, the English system profits from such supplementary data as the Wall Street Journal corpus, with which we have considerable prior experience.

- We used a phoneme set with toned versions of appropriate phonemes. This appeared to be a win when we were doing our development work, although follow-up analysis after the evaluation makes it appear less valuable in the final system.

- We started our Mandarin system from a flat start, while we used a small WSJ-trained recognizer to produce our initial segmentations for the English system.

- We did not use questions about the position of word boundaries in the decision tree state clustering. This was a significant win for us in our English system (1% absolute), but it did not prove to be so in Mandarin. This may be due in part to the fact that the notion of word, hence the position of a word boundary, is a somewhat arbitrary one.

- We did not use a phoneme recognizer in the automatic speech segmentation pass. In an early test, we tried using our English phoneme recognizer, reasoning that we were only looking for speech/silence distinctions. Unfortunately, the resulting system error rate was 5% points higher than expected, with almost all of this difference due to extra deletions, presumably caused by mistakenly labelling speech regions as silence/music. Therefore we chose to use a simple Mandarin word recognizer for the segmentation pass, which performed as well as we expected and wasn't too much slower than the English phone recognizer.

There are many other adjustments we might have made, most notably the addition of pitch to our front-end feature set. We look forward to trying out several of these in the future.

2. ACOUSTIC MODELS

The Mandarin system uses a 101-element phoneme set, including toned versions for all vowels and for the syllable-final glides /y/ and /w/. In most cases, this involved 5 tone variants (including the "neutral" fifth tone), but in a few cases no training data was available for some toned versions.

We built relatively small acoustic models, containing about 5000 states with up to 16 components. The models were trained entirely from the Mandarin acoustic training corpus.

This training corpus is much more homogeneous than the English training corpus. The 1996 official HUB4 (English) acoustic training corpus has about 33 hours of usable speech, making it similar in size to the 1997 Mandarin acoustic training corpus, which consists of about 26 hours of usable speech. But the corpora have quite different characteristics. The Mandarin data is gender balanced, while the English data is about 1/3 female and 2/3 male. The Mandarin data has 834 training speakers, while the English data has 1874 (far more than can be attributed to the size differences alone). The Mandarin data comes from only three broadcast sources, while the English data comes from 11 sources. This homogeneity of the Mandarin data makes for a far easier recognition task, which is reflected in the system error rates.

3. LANGUAGE MODELS

Three backoff trigram language models were trained from a total of about 100 million words of text.

- The Mandarin Broadcast News acoustic training transcriptions were used to train a language model (A) with 137,000 bigrams and 217,000 trigrams.

- The LDC Chinese Radio texts were used to train a language model (B) with 3.3 million bigrams and 3.5 million trigrams.

- The LDC People's Daily and Xinhua Newswire texts were used to train a model (C) with 10.4 million bigrams and 8.0 million trigrams.

These models were combined with interpolation weights 0.65 for A, 0.12 for B, and 0.23 for C, obtained by minimizing perplexity on the 1997 Hub4 Mandarin devtest. The three language models share a 45K vocabulary constructed by selecting all of the words from the language model training texts for which we had pronunciations. Pronunciations were drawn from the Mandarin03 lexicon (plus the HUB4 Supplement), provided by the LDC. We also added a small number of pronunciations, generated by hand, to cover anomalous entries such as word fragments in the training data. We only included English words if they occurred in the acoustic training texts.

Only the Mandarin Broadcast News acoustic training transcripts were made available with word-level tokenization. This tokenization had been performed by the transcribers. The other text sources had to be tokenized automatically, which we did by using a dynamic programming algorithm that had been developed as part of our work on the CallHome Mandarin corpus [2].

4. EXPERIMENTS & ANALYSIS

We first look at the performance of an increasingly sophisticated series of acoustic models. In Table 1, our baseline acoustic models are gender independent (GI) with no tones. To these basic models, we then add speaker normalization (SN). Then to the SN models we add toned phonemes, expanding the phoneme set from 39 to 101 elements (SN-Toned). Lastly, we add to the SN-Toned system the ability for the decision trees to entertain questions about the position of word boundary (WordBndry).

The test set is the 1997 HUB4 Mandarin devtest, segmented by using the hand-labelled turn marks and clustered by using the known speaker identities. For the speaker-normalized models, warp scales are chosen on a per cluster basis, and warped test data is used only with warped models. There is no adaptation.

The first two columns of Table 1 present the word error rate and character error rate using a 17k lexicon and trigram language model trained from the Mandarin acoustic training transcripts. The last column presents the character error rate using the 45k lexicon and interpolated trigram language model used for the 1997 evaluation.

The apparent value of the various acoustic improvements seems to depend on the strength of the language model used in testing. For example, with the weaker model tone appears to be a valuable addition to the system. However, with the stronger

System	WER	CER	Interp CER
GI	38.2	26.7	21.4
SN	37.2	26.1	19.7
SN-Toned	35.9	24.9	19.7
WordBndry	36.3	---	---

Table 1: Performance of increasingly sophisticated acoustic models with 2 language models. The first two columns use a simple 17k LM and the third column an interpolated 45k LM. (WER = word error rate; CER = character error rate)

language model this improvement disappears, perhaps because the language model is already able to fix near-homophone confusions that we had counted on the tone variants to distinguish. In fact, many of the errors being made by all of these systems involve short two-character homophones, and the more powerful language models are able to more reliably fix these errors. Unfortunately, the Mandarin devtest is relatively small (only about 53 minutes of speech), and it would be interesting to evaluate these improvements on a larger test set.

After the evaluation, we decided to explore alternatives to the toned system used in the evaluation. The test set and protocols in the following experiments are the same as in the experiments reported above. We used the 17k trigram LM for our initial experiments, and speaker normalized models and test data.

The motivation behind these experiments is that our toned system includes 101 phonemes, which may be too many distinctions to be modelled reliably by such a small corpus of training data. So we tried various schemes to overcome this deficiency.

The first thing we tried was to separately model only phenomena associated with the third tone. We kept third tone distinctions, but mapped all other tones to a common "toneless" phoneme -- see the 'tone3' entry in Table 2. Next we kept the third and fourth tone distinctions only ('tone34'). Since these experiments led to no improvement over our baseline toned system ('allTones' = 'SN-Toned' from Table 1), we decided to try using phoneme groups with our 101-element toned system. In this scheme a shared decision tree is grown for all toned variants of a given phoneme. Our state-clustering decision trees now support the ability to grow a single tree for a family of phonemes, where phonemes in the group share a common root node, and questions may be asked about the identity of the central phoneme as the tree is grown, in addition to the usual questions about phonetic context. In this way, we let the training data decide which tones should be modeled. This led to a slight improvement, but one which may not be statistically significant. The 'PG15' entry in Table 2 corresponds to the system with 101 phonemes, in which the tone variants were clustered into 15 groups.

Of course, more work needs to be done in our analysis of tone, especially since we have not yet included a pitch feature in our HUB4 front-end. In our 1997 CallHome Mandarin system [2],

we saw a small but significant improvement when adding toned phonemes, with a somewhat larger improvement after we added a pitch feature.

	#Phonemes	CER
noTones	39	26.1
tone3	58	25.2
tone34	73	24.9
allTones	101	24.9
PG15	101 (w/ 15grps)	24.5

Table 2: Experiments with tone. (CER gives character error rate on the Mandarin 1997 devtest.)

We were suspicious about the toned system improvements vanishing when we moved to a more powerful language model, as demonstrated in Table 1. We entertained the possibility that the relative weights of the language model and the acoustic models were miss-tuned so we ran a number of tuning experiments on the devtest. These experiments convinced us that this was not the case. However, in the tuning process we noticed that when we opened the beam width wider, we began to see significant decreases in the error rate. We re-ran the 1997 evaluation with the best settings obtained on the devtest, but leaving all other aspects of the system unchanged. The result is given in Table 3.

	CER
official result	20.2
after tuning	19.3

Table 3: Re-running the evaluation system with a wider beam width.

5. FUTURE WORK

The development that we undertook for the 1997 evaluation involved simply taking the English system's (word-based) architecture and getting it to work on the Mandarin data. However, it would be extremely interesting to develop character- or syllable-based recognizers for this task. We also plan to explore several possibilities for enhancing Mandarin recognition, such as the addition of pitch to our front-end feature set.

Acknowledgements

The HUB4 Mandarin system profited enormously from lessons learned in creating Dragon's Mandarin CallHome recognizer. We particularly wish to thank Yoshiko Ito for her generosity in helping us transfer knowledge from the CallHome system. We also thank Mark Mandel and Shen-Yi Luo for their invaluable assistance in developing the lexicon and in answering innumerable Mandarin language questions.

This work was supported by the Defense Advanced Research Projects Agency. The views and findings contained in this material are those of the authors and do not necessarily reflect the position or policy of the U. S. Government and no official endorsement should be inferred.

REFERENCES

[1] S. Wegmann et al., "Dragon Systems' 1997 Broadcast News Transcription System," *these Proceedings*.

[2] Y. Ito et al., "Dragon Systems' 1997 Mandarin CallHome Evaluation System," *Proc. HUB-5 Conversational Speech Recognition Workshop*, MITAGS, November 1997.